RECOVERING NORMAL NETWORKS FROM SHORTEST INTER-TAXA DISTANCE INFORMATION

MAGNUS BORDEWICH, KATHARINA T. HUBER, VINCENT MOULTON, AND CHARLES SEMPLE

Abstract. Phylogenetic networks are a type of leaf-labelled, acyclic, directed graph used by biologists to represent the evolutionary history of species whose past includes reticulation events. A phylogenetic network is tree-child if each non-leaf vertex is the parent of a tree vertex or a leaf. Up to a certain equivalence, it has been recently shown that, under two different types of weightings, edge-weighted tree-child networks are determined by their collection of distances between each pair of taxa. However, the size of these collections can be exponential in the size of the taxa set. In this paper, we show that, if we have no "shortcuts", that is, the networks are normal, the same results are obtained with only a quadratic number of inter-taxa distances by using the shortest distance between each pair of taxa. The proofs are constructive and give cubic-time algorithms in the size of the taxa sets for building such weighted networks.

1. Introduction

Distance-based methods collectively provide fundamental tools for the reconstruction and analysis of phylogenetic (evolutionary) trees. Two of the most popular and longstanding distance-based methods are UPGMA [14] and Neighbor Joining [11]. Loosely speaking, these methods take as input a distance matrix $\mathcal{D}$ on a set X of taxa, whose entries are the distances between pairs of taxa in X, and return an edge-weighted phylogenetic tree on X that best represents $\mathcal{D}$. Distances between taxa could, for example, measure the time since the two taxa separated from a common ancestor. Typically, the property underlying any distance-based method is the following: if $\mathcal{T}$ is a phylogenetic tree whose edges are assigned a positive real-valued weight, then the pairwise distances between taxa is sufficient to determine $\mathcal{T}$ and

Date: January 15, 2018.

1991 *Mathematics Subject Classification.* 05C85, 68R10.

Key words and phrases. Distance matrix, tree-child network, normal network.

The second and third authors were supported by the London Mathematical Society. The fourth author was supported by the New Zealand Marsden Fund.

its edge weighting [6, 13, 20]. This property is frequently referred to as the Tree-Metric Theorem (see [12, Theorem 7.2.6]).

While there exist numerous distance-based methods for reconstructing and analysing phylogenetic trees that aim to explicitly represent the treelike evolution of species, there are only a small number of such methods for phylogenetic networks. Yet, phylogenetic networks are necessary to accurately represent the ancestral history of sets of taxa whose evolution includes non-treelike (reticulate) evolutionary processes such as hybridisation and lateral gene transfer. As a step towards developing practical distance-based methods for reconstructing and analysing phylogenetic networks, in this paper we are interested in establishing analogues of the Tree-Metric Theorem for edge-weighted phylogenetic networks. In particular, we establish two such analogues for the increasingly prominent class of tree-child networks. Briefly, we show that, under two types of weightings, edge-weighted tree-child networks on X with no shortcuts can be reconstructed from the pairwise shortest distances between taxa in time polynomial in the size of X. The first type of weighting induces an ultrametric, while the second type of weighting has the property that the pair of edges directed into a reticulation (*i.e.* a vertex of in-degree two) have equal weight. We envisage that these results could lead to practical algorithms to construct phylogenetic networks from distance data.

The work in this paper is not the first to consider constructing phylogenetic networks from distances. Chan *et al.* [8] take a matrix of inter-taxa distances and reconstruct an ultrametric galled network having the property that there is a path between each pair of taxa having the same length as that given in the matrix, if such a network exists. Willson [18] studied the problem of determining a phylogenetic network given the average distance between each pair of taxa, where each reticulation assigns a probability to the two edges directed into it. It is shown in [18] that such distances are enough to determine a phylogenetic network with a single reticulation cycle in polynomial time. Bordewich and Semple [3], the original starting point for this paper, showed that (unweighted) tree-child networks can be reconstructed from the multi-set of distances between taxa in polynomial time. Other methods have been developed for building phylogenetic networks from distance data (see, for example, [16, 19]) but these use different approaches to associate a distance to a network than the ones presented here. In addition, Huber and Scholz [9] considered the problem of reconstructing phylogenetic networks from so-called symbolic distances, and it would be interesting to see if our new results could extend to such distances. Two other papers (specificially [4, 5]) that are particularly relevant to the work of our paper, are discussed in more detail in the next section.

Although distance based methods may be considered not as accurate as other methods such as Maximum Likelihood, they still have an important role in studying large datasets and gaining quick results when exploring data. This role may even be more significant for phylogenetic networks, where the complexity of inferring optimal solutions is even higher than for phylogenetic trees. The next section details some background and gives the statements of the two main results.

2. Main Results

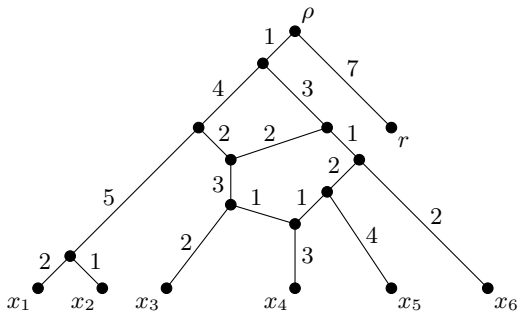
Throughout the paper, X denotes a non-empty finite set. A *phylogenetic network $\mathcal{N}$ on X* is a rooted acyclic directed graph with no parallel edges satisfying the following:

(i) the unique root has out-degree two;
(ii) the set X is the set of vertices of out-degree zero, each of which has in-degree one; and
(iii) all other vertices either have in-degree one and out-degree two, or in-degree two and out-degree one.

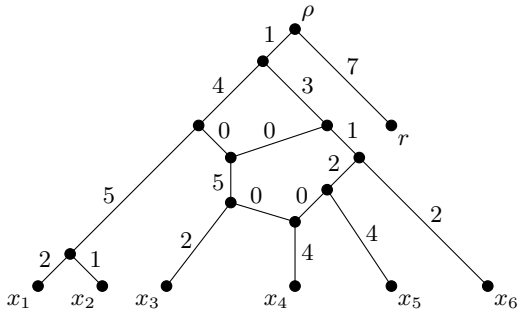
If $|X| = 1$, then we additionally allow the directed graph consisting of the single vertex in X to be a phylogenetic network. A phylogenetic network, as defined here, is often referred to as a binary phylogenetic network. The vertices in X are called *leaves*, while the vertices of in-degree one and out-degree two are *tree vertices*, and the vertices of in-degree two and out-degree one are *reticulations*. An edge directed into a reticulation is a *reticulation edge*; all other edges are *tree edges*. An element in X is an *outgroup* if its parent is the root of $\mathcal{N}$. Furthermore, an edge $e = (u, v)$ is a *shortcut* if there is a directed path in $\mathcal{N}$ from u to v avoiding e. Note that, necessarily, a shortcut is a reticulation edge. In the literature, shortcuts are also known as *redundant* edges. Ignoring the weighting of the edges, Fig. 1(i) shows a phylogenetic network with root ρ, outgroup r, and $X = \{r, x_1, x_2, x_3, x_4, x_5, x_6\}$. It has exactly two reticulations, but no shortcuts.

Let $\mathcal{N}$ be a phylogenetic network. We say $\mathcal{N}$ is *tree-child* if each non-leaf vertex in $\mathcal{N}$ is the parent of either a tree vertex or a leaf. Moreover, $\mathcal{N}$ is *normal* if it is tree-child and has no shortcuts. For example, the phylogenetic network in Fig. 1(i) is normal. As with all figures in this paper, edges are directed down the page. Tree-child networks and normal networks were introduced by Cardona et al. [7] and Willson [15, 17], respectively.

Let $\mathcal{N}$ be a phylogenetic network on X, and let E denote the edge set of $\mathcal{N}$. A *weighting* of $\mathcal{N}$ is a function $w : E \to \mathbb{R}$ that assigns edges a non-negative real-valued weight such that tree edges are assigned a strictly



(i) $(\mathcal{N}, w)$



(ii) $(\mathcal{N}, w')$

FIGURE 1. Two (weighted) phylogenetic networks on $\{r, x_1, x_2, x_3, x_4, x_5, x_6\}$.

positive weight. Thus reticulation edges are allowed weight zero. Let $(\mathcal{N}, w)$ be a weighted phylogenetic network, and let v and v' be vertices in $\mathcal{N}$. An *up-down path* P from v to v' is an underlying path

$$v = u_0, u_1, u_2, \ldots, u_k = v',$$

where, for some i,

$$(u_i, u_{i-1}), (u_{i-1}, u_{i-2}), \ldots, (u_1, v)$$

and

$$(u_i, u_{i+1}), (u_{i+1}, u_{i+2}), \ldots, (u_{k-1}, v')$$

are edges in $\mathcal{N}$. The *length* of P is the sum of the weights of the edges in P and, provided v and v' are distinct, u_i is the *peak* of P. In Fig. 1(i), there are exactly two up-down paths joining x_3 and x_5. One path has weight 21, while the other path has weight 14. As in this example, in this paper we are

interested in the *inter-taxa distances* in $(\mathcal{N}, w)$, that is, the lengths of the up-down paths connecting leaves in X. We next state two recent results [5, 4]. Both results prompted the two main results in this paper.

Let $(\mathcal{N}, w)$ be a weighted phylogenetic network on X and, for all $x, y \in X$, let $\mathcal{P}_{x,y}$ be the set of up-down paths from x to y in $(\mathcal{N}, w)$. The *multi-set of distances from x to y*, denoted $\mathcal{D}_{x,y}$, is the multi-set of the lengths of the paths in $\mathcal{P}_{x,y}$. Similarly, the *set of distances from x to y*, denoted $\overline{\mathcal{D}}_{x,y}$, is the set of the lengths of the paths in $\mathcal{P}_{x,y}$. Note that $\mathcal{D}_{x,y} = \mathcal{D}_{y,x}$, $\mathcal{D}_{x,x} = \{0\}$, $\overline{\mathcal{D}}_{x,y} = \overline{\mathcal{D}}_{y,x}$, and $\overline{\mathcal{D}}_{x,x} = \{0\}$ for all $x, y \in X$. The *multi-set distance matrix $\mathcal{D}$ of* $(\mathcal{N}, w)$ is the $|X|\times|X|$ matrix whose (x, y)-th entry is $\mathcal{D}_{x,y}$ (respectively, the *set distance matrix $\overline{\mathcal{D}}$ of* $(\mathcal{N}, w)$ is the $|X| \times |X|$ matrix whose (x, y)-th entry is $\overline{\mathcal{D}}_{x,y}$) for all $x, y \in X$, in which case we say $\mathcal{D}$ (resp. $\overline{\mathcal{D}}$) is *realised by* $(\mathcal{N}, w)$.

Let $(\mathcal{N}, w)$ be a weighted phylogenetic network on X and let $\mathcal{D}$ be the multi-set distance matrix of $(\mathcal{N}, w)$. The underlying problem we are investigating is determining how much information $\mathcal{D}$ contains about $(\mathcal{N}, w)$. The following highlights that, even with tree edges having positive weights, the weighting of $(\mathcal{N}, w)$ cannot be determined exactly. Let u be a reticulation in $\mathcal{N}$ with parents p_u and q_u, and let v be the unique child of u. We can change the weighting of the edges incident with u without changing $\mathcal{D}$. In particular, provided the sum of the weights of (p_u, u) and (u, v) equal

$$w(p_u, u) + w(u, v)$$

and the sum of the weights of (q_u, u) and (u, v) equal

$$w(q_u, u) + w(u, v),$$

we can change the weights of (p_u, u), (q_u, u), and (u, v) (keeping the weights of all other edges the same) to construct a different weighting, w' say, so that $\mathcal{D}$ is realised by $(\mathcal{N}, w')$. We refer to this change as *re-weighting the edges at a reticulation* of $\mathcal{N}$. For example, in Fig. 1, $(\mathcal{N}, w')$ has been obtained from $(\mathcal{N}, w)$ by reweighting the edges at both reticulations. If $\mathcal{N}$ has an outgroup r, a similar occurrence happens with the two edges incident with the root ρ of $\mathcal{N}$. In particular, now let u be the child of ρ that is not r. Note that u is a tree vertex. Then, provided the weights of (ρ, r) and (ρ, u) sum to

$$w(\rho, r) + w(\rho, u),$$

we can change the weights of (ρ, r) and (ρ, u) (keeping the weights of all other edges the same) to construct a different weighting, w'' say, so that $\mathcal{D}$ is realised by $(\mathcal{N}, w'')$. We refer to this change as *re-weighting the edges at the root* of $\mathcal{N}$.

Let $(\mathcal{N}, w)$ and $(\mathcal{N}_1, w_1)$ be two weighted phylogenetic networks on X. We say $(\mathcal{N}, w)$ and $(\mathcal{N}_1, w_1)$ are *equivalent* if $\mathcal{N}$ is isomorphic to $\mathcal{N}_1$, and w_1 can be obtained from w by re-weighting the edges at each reticulation and

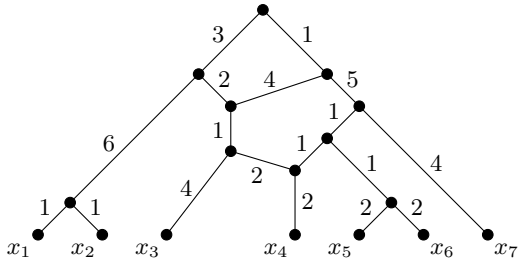


FIGURE 2. A phylogenetic network with an equidistant weighting.

re-weighting the edges at the root. For example, the weighted phylogenetic networks in Fig. 1 are equivalent.

We now state the two recent results. A weighting w of a phylogenetic network $\mathcal{N}$ is *equidistant* if the length of all paths starting at the root and ending at a leaf are the same length. A weighting w of a phylogenetic network is a *reticulation-pair weighting* if, for each reticulation v in $(\mathcal{N}, w)$, the two edges directed into v have the same weight. The weightings of the phylogenetic networks shown in Fig. 1 are reticulation-pair weightings. The weighting of the phylogenetic network shown in Fig. 2 is an equidistant weighting. The following theorems are established in [5] and [4], respectively.

Theorem 2.1. *Let $\overline{\mathcal{D}}$ be the set distance matrix of an equidistant-weighted tree-child network $(\mathcal{N}, w)$ on X. Then, up to equivalence, $(\mathcal{N}, w)$ is the unique such network realising $\overline{\mathcal{D}}$, in which case a member of its equivalence class can be found from $\overline{\mathcal{D}}$ in $O(|X|^4)$ time.*

Theorem 2.2. *Let $\mathcal{D}$ be the multi-set distance matrix with distinguished element r of a reticulation-pair weighted tree-child network $(\mathcal{N}, w)$ on X with outgroup r. Then, up to equivalence, $(\mathcal{N}, w)$ is the unique such network realising $\mathcal{D}$, in which case a member of its equivalence class can be found from $\mathcal{D}$ in time $O(|\mathcal{D}|^2)$.*

Theorems 2.1 and 2.2 are somewhat surprising given that, in the size of the leaf set, it is possible to have exponentially-many up-down paths connecting leaves in a tree-child network. For example, Fig. 3 shows a normal network with $2n+1$ leaves $x_1, x_2, \ldots, x_{2n+1}$ in which there are 2^n distinct up-down paths connecting x_1 and x_{2n+1}. Nevertheless, the fact that all such paths are considered is not satisfactory. The next two theorems are the two main results of this paper. They show that, up to shortcuts, we can retain the outcomes of Theorems 2.1 and 2.2 by knowing only a quadratic number of inter-taxa distances.

Let $(\mathcal{N}, w)$ be a weighted phylogenetic network on X. The *minimum distance matrix $\mathcal{D}_{\min}$ of* $(\mathcal{N}, w)$ is the $|X| \times |X|$ matrix whose (x, y)-th entry

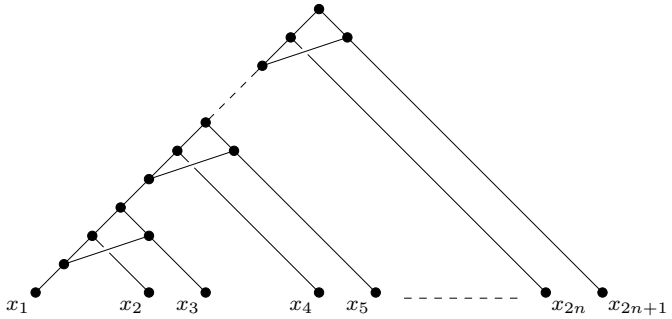


FIGURE 3. A normal network on $\{x_1, x_2, \ldots, x_{2n+1}\}$. There are 2^n distinct up-down paths connecting x_1 and x_{2n+1}.

is the minimum length of an up-down path joining x and y for all $x, y \in X$. We denote the (x, y)-th entry in $\mathcal{D}_{\min}$ by $d_{\min}(x, y)$.

Theorem 2.3. *Let $\mathcal{D}_{\min}$ be the minimum distance matrix of an equidistant-weighted normal network $(\mathcal{N}, w)$ on X. Then, up to equivalence, $(\mathcal{N}, w)$ is the unique such network realising $\mathcal{D}_{\min}$, in which case a member of its equivalence class can be found from $\mathcal{D}_{\min}$ in $O(|X|^3)$ time.*

Now let $(\mathcal{N}, w)$ be a weighted phylogenetic network on $X \cup \{r\}$, where r is an outgroup. For the purposes of the next theorem, the minimum distance matrix $\mathcal{D}_{\min}$ of $(\mathcal{N}, w)$ is the $|X| \times |X|$ matrix whose (x, y)-th entry is $d_{\min}(x, y)$ for all $x, y \in X$. Furthermore, the *maximum distance outgroup vector*, denoted $\mathbf{d}_{\max}$, is the vector of length $|X|$ whose x-th coordinate is the maximum length of an up-down path joining r and x for all $x \in X$. We denote the x-th coordinate in $\mathbf{d}_{\max}$ by $d_{\max}(r, x)$. The distances in $\mathbf{d}_{\max}$ are necessary in the way we establish the next theorem.

Theorem 2.4. *Let $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ be the minimum distance matrix and maximum distance outgroup vector of a reticulation-pair weighted normal network $(\mathcal{N}, w)$ on $X \cup \{r\}$, where r is an outgroup. Then, up to equivalence, $(\mathcal{N}, w)$ is the unique such network realising $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$, in which case a member of its equivalence class can be found from $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ in $O(|X|^3)$ time.*

It is easily seen that it is not possible to extend Theorems 2.3 and 2.4 to tree-child networks as the distance information given in the hypothesis of these theorems is insufficient to determine shortcuts. However, these results do hold for temporal tree-child networks as such networks have no shortcuts and are therefore normal. For the definition of temporal phylogenetic network, see [1].

The rest of the paper is organised as follows. The next section contains some necessary preliminaries. In particular, it contains the constructions

of the three operations that underlie the inductive proofs of Theorems 2.3 and 2.4. These proofs, including explicit descriptions of the associated algorithms, are given in Sections 4 and 5, respectively. The last section, Section 6, contains a brief conclusion.

3. Preliminaries

Let $\mathcal{N}$ be a phylogenetic network on X and let $\{s,t\}$ be a 2-element subset of X. Denote the unique parents of s and t by p_s and p_t, respectively. We call $\{s,t\}$ a *cherry* if $p_s = p_t$, that is, the parents of s and t are the same. Furthermore, we call $\{s,t\}$ a *reticulated cherry* if either p_s or p_t, say p_t, is a reticulation and p_s is a parent of p_t, in which case t is the *reticulation leaf* of the reticulated cherry. Observe that p_s is necessarily a tree vertex. To illustrate, in Fig. 1(i), $\{x_1, x_2\}$ is a cherry, while $\{x_3, x_4\}$ is a reticulated cherry in which x_4 is the reticulation leaf. In the same figure, $\{x_4, x_5\}$ is also a reticulated cherry. The next lemma is well known for tree-child networks (for example, see [3]). The restriction to normal networks is immediate. We will used it freely throughout the paper.

Lemma 3.1. *Let $\mathcal{N}$ be a normal network on X. Then*

(i) *If $|X| = 1$, then $\mathcal{N}$ consists of the single vertex in X, while if $|X| = 2$, say $X = \{s,t\}$, then $\mathcal{N}$ consists of the cherry $\{s,t\}$.*
(ii) *If $|X| \geq 2$, then $\mathcal{N}$ has either a cherry or a reticulated cherry.*

In addition to using the last lemma freely, we will also use freely the following observation. If $\mathcal{N}$ is a normal network and u is a vertex of $\mathcal{N}$, then there is a directed path from u to a leaf avoiding reticulations except perhaps u.

Let $(\mathcal{N}, w)$ be a weighted phylogenetic network on X. We now describe three operations on $(\mathcal{N}, w)$. The first and second operations underlie the inductive approach we take to prove Theorem 2.3, while the first and third operations underlie the inductive approach we take to prove Theorem 2.4. Let $\{s,t\}$ be 2-element subset of X, and denote the parents of s and t by p_s and p_t, respectively. First suppose that $\{s,t\}$ is a cherry. Let g_s denote the parent of p_s. Then *reducing* t is the operation of deleting t and its incident edge, suppressing p_s, and setting the weight of the resulting edge (g_s, s) to be

$$w(g_s, p_s) + w(p_s, s).$$

Now suppose that $\{s,t\}$ is a reticulated cherry, in which t is the reticulation leaf. Let g_s denote the parent of p_s, and let g_t denote the parent of p_t that is not p_s. Then *cutting* $\{s,t\}$ is the operation of deleting (p_s, p_t), suppressing p_s and p_t, and setting the weight of the resulting edge (g_s, s) to be

$$w(g_s, p_s) + w(p_s, s)$$

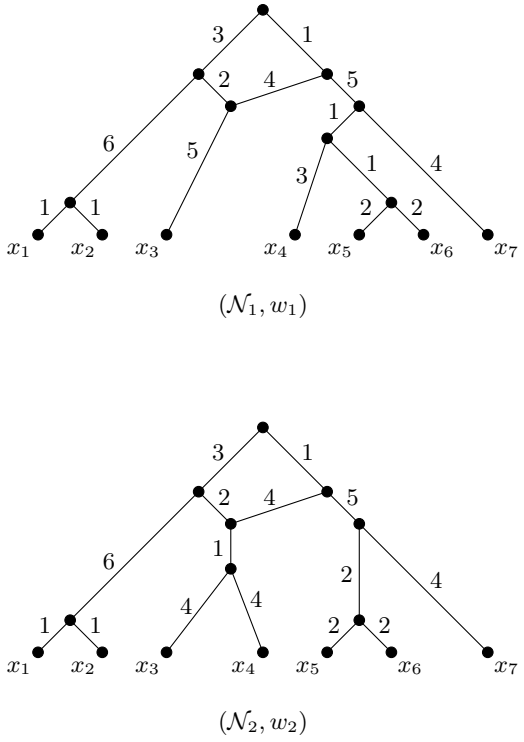


FIGURE 4. The weighted phylogenetic networks $(\mathcal{N}_1, w_1)$ and $(\mathcal{N}_2, w_2)$ obtained from the weighted phylogenetic network in Fig. 2 by cutting $\{x_3, x_4\}$ and isolating $\{x_3, x_4\}$, respectively.

and the edge (g_t, t) to be

$$w(g_t, p_t) + w(p_t, t).$$

Lastly, if g_t is a tree vertex and a parent of a tree vertex or a leaf, then *isolating* $\{s, t\}$ is the operation of deleting (g_t, p_t), suppressing g_t and p_t, and setting the weight of the resulting edge (p_s, t) to be

$$w(p_s, p_t) + w(p_t, t)$$

and the edge (g'_t, h) to be

$$w(g'_t, g_t) + w(g_t, h),$$

where g'_t is the parent of g_t and h is the child of g_t that is not p_t. To illustrate the last two operations, consider Fig. 4. The weighted phylogenetic network $(\mathcal{N}_1, w_1)$ has been obtained from the weighted phylogenetic network in Fig. 2 by cutting $\{x_3, x_4\}$, while $(\mathcal{N}_2, w_2)$ has been obtained from the same network by isolating $\{x_3, x_4\}$.

The proof of the next lemma is straightforward and omitted.

Lemma 3.2. *Let $(\mathcal{N}, w)$ be a weighted normal network. Let $\{s,t\}$ be a cherry or a reticulated cherry of $\mathcal{N}$. If $(\mathcal{N}', w')$ is obtained from $(\mathcal{N}, w)$ by reducing t if $\{s,t\}$ is a cherry, or by cutting or isolating $\{s,t\}$ if $\{s,t\}$ is a reticulated cherry, then $\mathcal{N}'$ is a normal network. Furthermore, if w is an equidistant (resp. reticulation-pair) weighting, then w' is an equidistant (resp. reticulation-pair) weighting.*

We next describe three operations on distance matrices that parallel the operations of reducing, cutting, and isolating. Let $\mathcal{D}$ be a distance matrix on X with each entry consisting of a single value, that is, $\mathcal{D}$ is an $|X| \times |X|$ matrix whose (x,y)-th entry is denoted $d(x,y)$. Let $\{s,t\}$ be a 2-element subset of X.

The first operation will be used only in association with reducing t when $\{s,t\}$ is a cherry. Let $\mathcal{D}'$ be the distance matrix on $X' = X - \{t\}$ obtained from $\mathcal{D}$ by setting

$$d'(x,y) = d'(y,x) = d(x,y)$$

for all $x, y \in X'$. We say that $\mathcal{D}'$ has been obtained from $\mathcal{D}$ by *reducing t in $\mathcal{D}$*.

The second operation will be used only in association with cutting $\{s,t\}$ when $\{s,t\}$ is a reticulation cherry in which t is the reticulation leaf, and the weighting is equidistant. Let

$$X_t = \{x \in X - \{s,t\} : d(t,x) \neq d(s,x)\}$$

and let $\delta = \min\{d(t,x) : x \in X_t\}$. Furthermore, let

$$X_\delta = \{x \in X_t : d(t,x) = \delta\}.$$

Intuitively, X_t are those leaves whose shortest path to t does not go through the parent of s, and X_δ are those members of X_t at minimum distance from t. To illustrate, consider the normal network with equidistant weighting shown in Fig. 2. Here $\{x_3, x_4\}$ is a reticulated cherry in which x_4 is the reticulation leaf. In this instance, $X_{x_4} = \{x_5, x_6, x_7\}$ and $\delta = 6$, so $X_\delta = \{x_5, x_6\}$. Now let $\mathcal{D}'$ be the distance matrix on X obtained from $\mathcal{D}$ by setting

$$d'(x,y) = d'(y,x) = d(x,y)$$

for all $x, y \in X - \{t\}$, setting $d'(t,t) = 0$, and setting

$$d'(t,y) = d'(y,t) = \max\{d(x,y) : x \in X_\delta - \{y\}\}$$

if $\max\{d(x,y) : x \in X_\delta - \{y\}\} \geq \delta$, and

$$d'(t,y) = d'(y,t) = \delta$$

otherwise for all $y \in X - \{t\}$. We say that $\mathcal{D}'$ has been obtained from $\mathcal{D}$ by *cutting $\{s,t\}$ in $\mathcal{D}$*. Intuitively, elements of X_δ are being used as a proxy for t in determining the minimum distances from t to members of $X_t - X_\delta$ in $\mathcal{D}'$; the distance from t to any member of X_δ remains δ. Thus, continuing the

illustration, if the equidistant weighting is denoted by $\mathcal{D}$ and $\mathcal{D}'$ is obtained from $\mathcal{D}$ by cutting $\{x_3, x_4\}$, then

$$d'(x_4, y) = d(x_5, y) = d'(x_5, y).$$

for all $y \in \{x_1, x_2, x_3\}$.

The third operation will be used only in association with isolating $\{s, t\}$ when $\{s, t\}$ is a reticulated cherry in which t is the reticulation leaf, r is an outgroup, and the weighting is a reticulation-pair weighting. Let r be a distinguished element in X, and let

$$\gamma = d(r, t) - d(r, s).$$

Intuitively, γ is the difference in length of the edge from the parent of s to s and the path from the parent of s to t. Let $\mathcal{D}'$ be the distance matrix on X obtained from $\mathcal{D}$ by setting

$$d'(x, y) = d'(y, x) = d(x, y)$$

for all $x, y \in X - \{t\}$,

$$d'(t, x) = d'(x, t) = d(s, x) + \gamma$$

for all $x \in X - \{s, t\}$, and

$$d'(t, s) = d'(s, t) = d(t, s).$$

We say that $\mathcal{D}'$ has been obtained from $\mathcal{D}$ by *isolating* $\{s, t\}$ *in* $\mathcal{D}$. For an illustration, consider the normal network with reticulation-pair weighting in Fig. 1(i). Now $\{x_3, x_4\}$ is a reticulated cherry in which t is the reticulation leaf. Here

$$\gamma = d(r, x_4) - d(r, x_3) = 2.$$

Thus if the reticulation-pair weighting is denoted by $\mathcal{D}$ and $\mathcal{D}'$ is obtained from $\mathcal{D}$ by isolating $\{x_3, x_4\}$, then

$$d'(x_4, y) = d(x_3, y) + 2$$

for all $y \in \{x_1, x_2, x_5, x_6, x_7\}$.

4. Proof of Theorem 2.3

In this section, we establish Theorem 2.3. We begin with two lemmas.

Lemma 4.1. *Let $(\mathcal{N}, w)$ be a equidistant-weighted normal network on X, where $|X| \geq 2$. Let $\mathcal{D}_{\min}$ be the minimum distance matrix of $(\mathcal{N}, w)$, and let $\{s, t\}$ be a 2-element subset of X such that*

$$d_{\min}(s, t) = \min\{d_{\min}(x, y) : x, y \in X\}.$$

Then $\{s, t\}$ is either a cherry or a reticulated cherry of $(\mathcal{N}, w)$. Moreover,

(i) *The set* $\{s,t\}$ *is a cherry of* $(\mathcal{N},w)$ *if* $d_{\min}(s,x) = d_{\min}(t,x)$ *for all* $x \in X - \{s,t\}$.
(ii) *Otherwise,* $\{s,t\}$ *is a reticulated cherry of* $(\mathcal{N},w)$ *in which* t *is the reticulation leaf precisely if* $d_{\min}(s,x) > d_{\min}(t,x)$ *for some* $x \in X - \{s,t\}$.

Proof. Let p_s and p_t denote the parents of s and t, respectively. If p_s and p_t are both reticulations, then, as $\mathcal{N}$ is normal and w equidistant, it is easily seen that there is an element $x \in X - \{s,t\}$ such that either $d_{\min}(s,t) > d_{\min}(s,x)$ or $d_{\min}(s,t) > d_{\min}(t,x)$; a contradiction ($x$ is a descendant of a tree vertex on the shortest up-down path from s to t that is not the peak of that up-down path). Thus, either p_s or p_t is a tree vertex. Without loss of generality, we may assume that p_s is a tree vertex. Let u denote the child of p_s that is not s. If u is a tree vertex, then, as w is equidistant,

$$d_{\min}(s,t) > d_{\min}(a,b) \geq \min\{d_{\min}(x,y) : x,y \in X\},$$

where $\{a,b\}$ is a cherry or reticulated cherry and a,b are descendants of u; a contradiction. Therefore u is either a leaf or a reticulation. If u is a leaf, then, as w is equidistant, $u = t$, in which case, $\{s,t\}$ is a cherry. If u is a reticulation, then, as $\mathcal{N}$ is normal and

$$d_{\min}(s,t) = \min\{d_{\min}(x,y) : x,y \in X\},$$

it follows that the unique child of u is t, and so $\{s,t\}$ is a reticulated cherry. Since $\mathcal{N}$ has no shortcuts and w is equidistant, it is easily checked that $\{s,t\}$ is a reticulated cherry in which t is the reticulation leaf if and only if $d_{\min}(s,t) > d_{\min}(t,x)$ for some $x \in X - \{s,t\}$. The lemma immediately follows. □

Lemma 4.2. *Let* $(\mathcal{N},w)$ *be an equidistant-weighted normal network on* X*, where* $|X| \geq 2$*. Let* $\mathcal{D}_{\min}$ *be the minimum distance matrix of* $(\mathcal{N},w)$*, and let* $\{s,t\}$ *be a 2-element subset of* X *such that*

$$d_{\min}(s,t) = \min\{d_{\min}(x,y) : x,y \in X\}.$$

Then the following hold:

(i) *If* $\{s,t\}$ *is a cherry, then the distance matrix obtained from* $\mathcal{D}_{\min}$ *by reducing* t *is the minimum distance matrix* $\mathcal{D}'_{\min}$ *realised by the weighted network* $(\mathcal{N}',w')$ *obtained from* $(\mathcal{N},w)$ *by reducing* t.
(ii) *If* $\{s,t\}$ *is a reticulated cherry in which* t *is the reticulation leaf, then the distance matrix obtained from* $\mathcal{D}_{\min}$ *by cutting* $\{s,t\}$ *is the minimum distance matrix* $\mathcal{D}'_{\min}$ *realised by the weighted network* $(\mathcal{N}',w')$ *obtained from* $(\mathcal{N},w)$ *by cutting* $\{s,t\}$.

Proof. By Lemma 4.1, $\{s,t\}$ is either a cherry or a reticulated cherry. Furthermore, by the same lemma, if $\{s,t\}$ is a reticulated cherry, we may assume, without loss of generality, that t is the reticulation leaf. Also, by Lemma 3.2, $(\mathcal{N}', w')$ is an equidistant-weighted normal network regardless of which of the two stated ways it is obtained from $(\mathcal{N}, w)$. If $\{s,t\}$ is a cherry, then it is clear that the distance matrix obtained from $\mathcal{D}_{\min}$ by reducing t is the minimum distance matrix of $(\mathcal{N}', w')$. Therefore, suppose that $\{s,t\}$ is a reticulated cherry, in which case $(\mathcal{N}', w')$ is obtained from $(\mathcal{N}, w)$ by cutting $\{s,t\}$. Let $\mathcal{D}'$ be the distance matrix obtained from $\mathcal{D}_{\min}$ by cutting $\{s,t\}$. We will show that $\mathcal{D}'$ is the minimum distance matrix $\mathcal{D}'_{\min}$ of $(\mathcal{N}', w')$.

Let p_s and p_t denote the parents of s and t, respectively, in $(\mathcal{N}, w)$. Since the only up-down paths in $(\mathcal{N}, w)$ joining elements in X that traverse the edge (p_s, p_t) involve t, it follows that $d'(x,y) = d'_{\min}(x,y)$ for all $x, y \in X - \{t\}$. Thus, to complete the proof, it suffices to show that $d'(t,x) = d'_{\min}(t,x)$ for all $x \in X - \{t\}$.

Let g_t denote the parent of p_t that is not p_s. Since $\mathcal{N}$ has no shortcuts, g_t is not an ancestor of s. Therefore, as $\mathcal{N}$ is normal, there is a (directed) path from g_t to a leaf, ℓ say, containing no reticulations, where $\ell \notin \{s,t\}$. Note that, in what follows, we never determine ℓ but its existence underlies the rest of the proof.

Let

$$X_t = \{x \in X - \{s,t\} : d_{\min}(t,x) \neq d_{\min}(s,x)\}.$$

Thus, if $x \in X_t$, then every minimum length up-down path in $(\mathcal{N}, w)$ joining t and x must traverse the edge (g_t, p_t). Observe that $\ell \in X_t$ as the edge directed into g_t, which is a tree edge, has positive weight and every up-down path from s to ℓ traverses this edge and therefore g_t, so $d_{\min}(t,\ell) < d_{\min}(s,\ell)$. Now let

$$\delta = \min\{d_{\min}(t,x) : x \in X_t\}$$

and let X_δ denote the subset of X_t consisting of those elements x such that $d_{\min}(t,x) = \delta$, that is,

$$X_\delta = \{x \in X_t : d_{\min}(t,x) = \delta\}.$$

Observe that, as w is equidistant, the lengths of all directed paths from g_t to a leaf are the same, and so $\ell \in X_\delta$. Therefore the elements in X_δ are descendants of g_t.

Let $y \in X - \{t\}$. We next determine whether or not y is a descendant of g_t. First note that $|X_\delta| = 1$ if and only if g_t is the parent of a leaf, in which case ℓ is the only leaf apart from t that is a descendant of g_t. So assume $|X_\delta| \geq 2$. We establish two claims:

(i) If
$$\max\{d_{\min}(x,y) : x \in X_\delta - \{y\}\} \geq \delta,$$
then y is not a descendant of g_t.

(ii) If
$$\max\{d_{\min}(x,y) : x \in X_\delta - \{y\}\} < \delta,$$
then y is a descendant of g_t.

To see (i) and (ii), if y is a descendant of g_t, then $d_{\min}(x,y) < \delta$ for all $x \in X_\delta - \{y\}$, so
$$\max\{d_{\min}(x,y) : x \in X_\delta - \{y\}\} < \delta.$$
On the other hand, if y is not a descendant of g_t, then
$$\max\{d_{\min}(x,y) : x \in X_\delta - \{y\}\} \geq d_{\min}(\ell, y) \geq d_{\min}(t,\ell) = \delta.$$

It follows from (i) and (ii) that, for all $y \in X - \{t\}$, if y is not a descendant of g_t, then
$$d'_{\min}(t,y) = \max\{d_{\min}(x,y) : x \in X_\delta - \{y\}\},$$
otherwise
$$d'_{\min}(t,y) = \delta.$$
Hence $d'(t,x) = d'_{\min}(t,x)$ for all $x \in X - \{t\}$, thereby completing the proof of the lemma. □

We next prove the uniqueness part of Theorem 2.3 which we restate for convenience.

Theorem 2.3. Let $\mathcal{D}_{\min}$ be the minimum distance matrix of an equidistant-weighted normal network $(\mathcal{N}, w)$ on X. Then, up to equivalence, $(\mathcal{N}, w)$ is the unique such network realising $\mathcal{D}_{\min}$, in which case a member of its equivalence class can be found from $\mathcal{D}_{\min}$ in $O(|X|^3)$ time.

Proof of the uniqueness part of Theorem 2.3. The proof is by induction on the sum of the number n of leaves and the number k of reticulations in $(\mathcal{N}, w)$. If $n+k = 1$, then $n = 1$ and $k = 0$, and $(\mathcal{N}, w)$ consists of the single vertex in X, and so uniqueness holds. If $n + k = 2$, then, as $\mathcal{N}$ is normal, $n = 2$ and $k = 0$, and $(\mathcal{N}, w)$ consists of two leaves attached to the root. Again, uniqueness holds as w is equidistant and so the weights of the edges incident with the leaves is fixed. Now suppose that $n+k \geq 3$, so $n \geq 2$, and that the uniqueness holds for all equidistant-weighted normal networks for which the sum of the number of leaves and the number of reticulations is at most $n + k - 1$.

Let $\{s, t\}$ be a 2-element subset of X such that
$$d_{\min}(s,t) = \min\{d_{\min}(x,y) : x, y \in X\}.$$

By Lemma 4.1, $\{s,t\}$ is either a cherry or a reticulated cherry of $(\mathcal{N},w)$. If $\{s,t\}$ is a reticulated cherry, then, by the same lemma, we can determine from $\mathcal{D}_{\min}$ which of s and t is the reticulation leaf. Thus, without loss of generality, we may assume that t is the reticulation leaf. Depending on whether $\{s,t\}$ is a cherry or a reticulated cherry, let $(\mathcal{N}',w')$ and $\mathcal{D}'$ be the weighted network and distance matrix obtained from $(\mathcal{N},w)$ and $\mathcal{D}_{\min}$ by reducing t or cutting $\{s,t\}$, respectively. By Lemma 4.2, $\mathcal{D}'$ is the minimum distance matrix of $(\mathcal{N}',w')$. Since $(\mathcal{N}',w')$ either has $n-1$ leaves and k reticulations if $\{s,t\}$ is a cherry, or n leaves and $k-1$ reticulations if $\{s,t\}$ is a reticulated cherry, it follows by the induction assumption that, up to equivalence, $(\mathcal{N}',w')$ is the unique equidistant-weighted normal network with minimum distance matrix $\mathcal{D}'$.

Let $(\mathcal{N}_1,w_1)$ be an equidistant-weighted normal network on X with minimum distance matrix $\mathcal{D}_{\min}$. By Lemma 4.1, $\{s,t\}$ is either a cherry or a reticulated cherry in $(\mathcal{N}_1,w_1)$. Indeed, by the same lemma, $\{s,t\}$ is a cherry in $(\mathcal{N}_1,w_1)$ precisely if it is a cherry in $(\mathcal{N},w)$. First assume that $\{s,t\}$ is a cherry in $(\mathcal{N},w)$. Then $\{s,t\}$ is a cherry in $(\mathcal{N}_1,w_1)$. Let $(\mathcal{N}_1',w_1')$ be the equidistant-weighted normal network obtained from $(\mathcal{N}_1,w_1)$ by reducing t. By Lemma 4.2, $\mathcal{D}'$ is the minimum distance matrix of $(\mathcal{N}_1',w_1')$ and so, by the induction assumption, $(\mathcal{N}_1',w_1')$ and $(\mathcal{N}',w')$ are equivalent. Using this equivalence and considering $d_{\min}(s,t)$, it is easily seen that $(\mathcal{N}_1,w_1)$ and $(\mathcal{N},w)$ are equivalent.

Now assume that $\{s,t\}$ is a reticulated cherry in $(\mathcal{N},w)$. Then $\{s,t\}$ is a reticulated cherry in $(\mathcal{N}_1,w_1)$ where, by Lemma 4.1, t is a reticulation leaf. Let $(\mathcal{N}_1',w_1')$ be the equidistant-weighted normal network obtained from $(\mathcal{N}_1,w_1)$ by cutting $\{s,t\}$. Since $\mathcal{D}_{\min}$ is the minimum distance matrix of $(\mathcal{N}_1,w_1)$, it follows by Lemma 4.2 that $\mathcal{D}'$ is the minimum distance matrix of $(\mathcal{N}_1',w_1')$. Therefore, by the induction assumption, $(\mathcal{N}_1',w_1')$ and $(\mathcal{N}',w')$ are equivalent. By again considering $d_{\min}(s,t)$, it is now easily deduced that $(\mathcal{N}_1,w_1)$ and $(\mathcal{N},w)$ are equivalent. This completes the proof of the uniqueness part of the theorem. □

4.1. **The Algorithm.** Let $(\mathcal{N},w)$ be an equidistant-weighted normal network on X and let $\mathcal{D}_{\min}$ denote the minimum distance matrix of $(\mathcal{N},w)$. We next give an algorithm which takes as input X and $\mathcal{D}_{\min}$, and returns a weighted network $(\mathcal{N}_0,w_0)$ equivalent to $(\mathcal{N},w)$. Its correctness is essentially established in proving the uniqueness part of Theorem 2.3 and so a formal proof of this is omitted. However, its running time is given at the end of this section. Called Equidistant Normal, the algorithm works as follows.

1. If $|X|=1$, then return the weighted phylogenetic network consisting of the single vertex in X.

2. If $|X| = 2$, say $X = \{s, t\}$, then return the weighted phylogenetic network with exactly two leaves s and t adjoined to the root by edges each with weight $\frac{1}{2}d_{\min}(s, t)$.
3. Else, find a 2-element subset $\{s, t\}$ of X such that

$$d_{\min}(s, t) = \min\{d_{\min}(x, y) : x, y \in X\}.$$

   (a) If $d_{\min}(s, x) = d_{\min}(t, x)$ for all $x, y \in X$ (so $\{s, t\}$ is a cherry), then
      (i) Reduce t in $\mathcal{D}_{\min}$ to give the distance matrix $\mathcal{D}'$ on $X' = X - \{t\}$.
      (ii) Apply Equidistant Normal to input X' and $\mathcal{D}'$. Construct $(\mathcal{N}_0, w_0)$ from the returned weighted phylogenetic network $(\mathcal{N}'_0, w'_0)$ on X' as follows. If p'_s is the parent of s in $(\mathcal{N}'_0, w'_0)$, then subdivide (p'_s, s) with a new vertex p_s, adjoin a new leaf t to p_s via the new edge (p_s, t), and set

$$w_0(p_s, s) = w_0(p_s, t) = \tfrac{1}{2}d_{\min}(s, t)$$

      and $w_0(p'_s, p_s) = w'_0(p'_s, s) - \frac{1}{2}d_{\min}(s, t)$. Keeping all other edge weights the same as their counterparts in $(\mathcal{N}'_0, w'_0)$, return $(\mathcal{N}_0, w_0)$.

   (b) Else ($\{s, t\}$ is a reticulated cherry, in which case, t is the reticulation leaf if there exists an $x \in X - \{s, t\}$ such that $d_{\min}(s, x) > d_{\min}(t, x)$),
      (i) Cut $\{s, t\}$ in $\mathcal{D}_{\min}$ to give the distance matrix $\mathcal{D}'$ on X.
      (ii) Apply Equidistant Normal to input X and $\mathcal{D}'$. Construct $(\mathcal{N}_0, w_0)$ from the returned weighted phylogenetic network $(\mathcal{N}'_0, w'_0)$ on X as follows. If p'_s and p'_t denote the parents of s and t in $(\mathcal{N}'_0, w'_0)$, respectively, then subdivide (p'_s, s) and (p'_t, t) with new vertices p_s and p_t, respectively, adjoin p_s and p_t via the new edge (p_s, p_t), set $w_0(p_s, s) = \frac{1}{2}d_{\min}(s, t)$ and $w_0(p'_s, p_s) = w'_0(p'_s, s) - \frac{1}{2}d_{\min}(s, t)$, and, for some positive real value ω such that $\omega \leq \frac{1}{2}d_{\min}(s, t)$ and $\omega \leq w'_0(p'_t, t)$, set $w_0(p_t, t) = \omega$, $w_0(p_s, p_t) = \frac{1}{2}d_{\min}(s, t) - \omega$, and $w_0(p'_t, p_t) = w'_0(p'_t, t) - \omega$. Keeping all other edge weights the same as their counterparts in $(\mathcal{N}'_0, w'_0)$, return $(\mathcal{N}_0, w_0)$.

We now consider the running time of Equidistant Normal. The algorithm takes as input a set X and an $|X| \times |X|$ distance matrix $\mathcal{D}_{\min}$ whose entries are the minimum length of an up-down path joining elements in X of an equidistant-weighted normal network $(\mathcal{N}, w)$ on X. Unless $|X| \in \{1, 2\}$, in which case Equidistant Normal runs in constant time, each iteration starts by finding a 2-element subset $\{s, t\}$ of X such that

$$d_{\min}(s, t) = \min\{d_{\min}(x, y) : x, y \in X\}.$$

This takes $O(|X|^2)$ time. Once such a 2-element subset is found, we compute $\mathcal{D}'$. This computation is done in one of two ways depending on whether or

not

$$d_{\min}(s,x) = d_{\min}(t,x)$$

for all $x \in X - \{s,t\}$. If, for some x,

$$d_{\min}(s,x) \neq d_{\min}(t,x),$$

we need to additionally check which of $d_{\min}(s,x) < d_{\min}(t,x)$ and $d_{\min}(s,x) > d_{\min}(t,x)$ hold. Thus the determination of which way to compute $\mathcal{D}'$ can be done in $O(|X|)$ time. Regardless of the way, $\mathcal{D}'$ can be computed in $O(|X|)$ time. Once $(\mathcal{N}_0', w_0')$ is returned, it can be augmented to $(\mathcal{N}_0, w_0)$ in constant time. Hence the total time of each iteration is $O(|X|^2)$ time.

When we recurse, the distance matrix $\mathcal{D}'$ inputted to the recursive call is the minimum distance matrix of a normal network with either one less leaf or one less reticulation than a normal network for which $\mathcal{D}_{\min}$ is the minimum distance matrix. Since a normal network has at most $|X| - 2$ reticulations, it has $O(|X|)$ vertices in total [2] (also see [10]), and so the total number of iterations is at most $O(|X|)$. Thus Equidistant Normal completes in $O(|X|^3)$ time. This completes the proof of Theorem 2.3.

5. Proof of Theorem 2.4

In this section, we prove Theorem 2.4. We begin with two lemmas. Let $\mathcal{N}$ be a phylogenetic network, and suppose that $\{s,t\}$ is either a cherry or a reticulated cherry in which t is the reticulation leaf in $\mathcal{N}$. Noting that the parent of s has out-degree two and is therefore a tree vertex, we refer to the parent of s as the *tree vertex of* $\{s,t\}$. For the next lemma, the proof of (i) and (iii) are given in [4], while the proof of (ii) is similar to that of Lemma 4.1 and is omitted. For $|X| \geq 2$, let $(\mathcal{N}, w)$ be a weighted network on $X \cup \{r\}$, where r is an outgroup. For all $x, y \in X$, we denote the value

$$\tfrac{1}{2}\{d_{\max}(r,x) + d_{\max}(r,y) - d_{\min}(x,y)\}$$

by $\mathcal{Q}_r(x,y)$. This value is crucial in obtaining the results in [4].

Lemma 5.1. *Let $|X| \geq 2$, and let $(\mathcal{N}, w)$ be a weighted tree-child network on $X \cup \{r\}$, where r is an outgroup. Let $\mathcal{D}_{\min}$ be the minimum distance matrix and let $\mathbf{d}_{\max}$ be the maximum distance outgroup vector of $(\mathcal{N}, w)$. Let $\{s,t\}$ be a 2-element subset of X such that*

$$\mathcal{Q}_r(s,t) = \max\{\mathcal{Q}_r(x,y) : x, y \in X\}.$$

Then

(i) *$\{s,t\}$ is either a cherry or a reticulated cherry of $(\mathcal{N}, w)$.*

(ii) $\{s,t\}$ *is a cherry of* $(\mathcal{N},w)$ *if and only if*

$$d_{\min}(s,x)+d_{\max}(r,t)-d_{\max}(r,s)=d_{\min}(t,x)$$

for all $x\in X-\{s,t\}$. *Otherwise,* $\{s,t\}$ *is a reticulated cherry in which* t *is the reticulation leaf if*

$$d_{\min}(s,x)+d_{\max}(r,t)-d_{\max}(r,s)>d_{\min}(t,x)$$

for some $x\in X-\{s,t\}$.

(iii) *The length of the longest up-down path in* $(\mathcal{N},w)$ *starting at* r *and ending at the tree vertex of* $\{s,t\}$ *is* $\mathcal{Q}_r(s,t)$, *and* $d_{\max}(r,s)$ *and* $d_{\max}(r,t)$ *are each realised by up-down paths that include this tree vertex.*

Lemma 5.2. *Let* $|X|\geq 2$, *and let* $(\mathcal{N},w)$ *be a reticulation-pair weighted normal network on* $X\cup\{r\}$, *where* r *is an outgroup. Let* $\mathcal{D}_{\min}$ *and* $\mathbf{d}_{\max}$ *be the minimum distance matrix and maximum distance outgroup vector of* $(\mathcal{N},w)$, *respectively. Let* $\{s,t\}$ *be a* 2*-element subset of* X *such that*

$$\mathcal{Q}_r(s,t)=\max\{\mathcal{Q}_r(x,y):x,y\in X\}.$$

Then the following hold:

(i) *If* $\{s,t\}$ *is a cherry, then the distance matrix and distance vector obtained from* $\mathcal{D}_{\min}$ *and* $\mathbf{d}_{\max}$ *by reducing* t *are the minimum distance matrix* $\mathcal{D}'_{\min}$ *and maximum distance outgroup vector* $\mathbf{d}'_{\max}$ *realised by the weighted network* $(\mathcal{N}',w')$ *obtained from* $(\mathcal{N},w)$ *by reducing* t.

(ii) *If* $\{s,t\}$ *is a reticulated cherry in which* t *is the reticulation leaf, then the distance matrix and distance vector obtained from* $\mathcal{D}_{\min}$ *and* $\mathbf{d}_{\max}$ *by isolating* $\{s,t\}$ *are the minimum distance matrix* $\mathcal{D}'_{\min}$ *and maximum distance outgroup vector* $\mathbf{d}'_{\max}$ *realised by the weighted network* $(\mathcal{N}',w')$ *obtained from* $(\mathcal{N},w)$ *by isolating* $\{s,t\}$.

Proof. By Lemma 5.1, $\{s,t\}$ is either a cherry or a reticulated cherry of $(\mathcal{N},w)$. By the same lemma, if $\{s,t\}$ is a reticulated cherry, then we may assume, without loss of generality, that t is the reticulation leaf. Furthermore, it follows by Lemma 3.2 that $(\mathcal{N}',w')$ is a reticulation-pair normal network with outgroup r. If $\{s,t\}$ is a cherry, then it is clear that the lemma holds. Therefore, suppose that $\{s,t\}$ is a reticulated cherry, in which case $(\mathcal{N}',w')$ is obtained from $(\mathcal{N},w)$ by isolating $\{s,t\}$. Let $\mathcal{D}'$ and $\mathbf{d}'$ be the distance matrix and distance vector obtained from $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ by isolating $\{s,t\}$. We will show that $\mathcal{D}'$ and $\mathbf{d}'$ are the minimum distance matrix $\mathcal{D}'_{\min}$ and maximum distance outgroup vector $\mathbf{d}'_{\max}$ of $(\mathcal{N}',w')$.

Let p_s and p_t denote the parents of s and t, respectively, and let g_t denote the parent of p_t that is not p_s in $(\mathcal{N},w)$. Note that, as $\mathcal{N}$ is normal, g_t is a tree vertex and not an ancestor of p_s. Since the only up-down paths in $(\mathcal{N},w)$ joining elements in X which traverse (g_t,p_t) involve t, it follows that

to complete the proof, it suffices to show that $d'(t,x) = d'_{\min}(t,x)$ for all $x \in X - \{t\}$ and $d'(r,t) = d'_{\max}(r,t)$.

By Lemma 5.1(iii),

$$d'(r,t) = d'_{\max}(r,t).$$

Furthermore, let $\gamma = d_{\max}(r,t) - d_{\max}(r,s)$. Then, by Lemma 5.1, since isolating $\{s,t\}$ creates a cherry $\{s,t\}$ in $(\mathcal{N}', w')$,

$$d'_{\min}(t,x) = d'_{\min}(s,x) + \gamma$$

for all $x \in X - \{s,t\}$, so $d'(t,x) = d'_{\min}(t,x)$ for all $x \in X - \{s,t\}$. Lastly, as w is a reticulation-pair weighting, $w(g_t, p_t) = w(p_s, p_t)$, so $d'(t,s) = d'_{\min}(t,s)$. This completes the proof of the lemma. □

We next establish the uniqueness part of Theorem 2.4. For convenience, we restate this theorem.

Theorem 2.4. Let $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ be the minimum distance matrix and maximum distance outgroup vector of a reticulation-pair weighted normal network $(\mathcal{N}, w)$ on $X \cup \{r\}$, where r is an outgroup. Then, up to equivalence, $(\mathcal{N}, w)$ is the unique such network realising $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$, in which case a member of its equivalence class can be found from $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ in $O(|X|^3)$ time.

Proof of the uniqueness part of Theorem 2.4. The proof is by induction on the sum of the number n of leaves and the number k of reticulations in $(\mathcal{N}, w)$. If $n + k = 1$, then $n = 1$ and $k = 0$, and so $(\mathcal{N}, w)$ consists of the single vertex in X, in which case the uniqueness holds. If $n + k = 2$, then $n = 2$, $k = 0$, and $(\mathcal{N}, w)$ consists of two leaves attached to the root, one of which is the outgroup r. Again, the uniqueness holds. Now suppose that $n + k \geq 3$, so $n \geq 3$, and the uniqueness holds for all reticulation-pair weighted normal networks for which the sum of the number of leaves and the number of reticulations is at most $n + k - 1$.

Let $\{s,t\}$ be a 2-element subset of X such that

$$\mathcal{Q}_r(s,t) = \max\{\mathcal{Q}_r(x,y) : x, y \in X\}.$$

By Lemma 5.1, $\{s,t\}$ is either a cherry or a reticulated cherry. If $\{s,t\}$ is a reticulated cherry, then, by the same lemma, $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ determine whether s or t is the reticulation leaf. Thus, without loss of generality, we may assume that t is the reticulation leaf. Let $(\mathcal{N}', w')$, $\mathcal{D}'$, and $\mathbf{d}'$ be the weighted phylogenetic network, distance matrix, and distance vector obtained from $(\mathcal{N}, w)$, $\mathcal{D}_{\min}$, and $\mathbf{d}_{\max}$, respectively, by reducing t if $\{s,t\}$ is a cherry or isolating $\{s,t\}$ if $\{s,t\}$ is a reticulated cherry. Now $(\mathcal{N}', w')$ either has $n-1$ leaves and k reticulations, or n leaves and $k-1$ reticulations, and so, by Lemma 5.2 and the induction assumption, up to equivalence, $(\mathcal{N}', w')$ is

the unique reticulation-pair weighted normal network with outgroup r whose minimum distance matrix is $\mathcal{D}'$ and maximum distance outgroup vector is $\mathbf{d}'$.

Let $(\mathcal{N}_1, w_1)$ be a reticulation-pair weighted normal network on X with outgroup r whose minimum distance matrix is $\mathcal{D}_{\min}$ and maximum distance outgroup vector is $\mathbf{d}_{\max}$. By Lemma 5.1, $\{s,t\}$ is either a cherry or a reticulated cherry in $(\mathcal{N}_1, w_1)$. Moreover, by the same lemma, $\{s,t\}$ is a cherry in $(\mathcal{N}_1, w_1)$ if and only if it is a cherry in $(\mathcal{N}, w)$. First assume that $\{s,t\}$ is a cherry in $(\mathcal{N}, w)$. Let $(\mathcal{N}_1', w_1')$ be the reticulation-pair weighted normal network with outgroup r obtained from $(\mathcal{N}_1, w_1)$ by reducing t. By Lemma 5.2, $\mathcal{D}'$ and $\mathbf{d}'$ are the minimum distance matrix and maximum distance outgroup vector of $(\mathcal{N}_1', w_1')$. Thus, by the induction assumption, $(\mathcal{N}_1', w_1')$ and $(\mathcal{N}', w')$ are equivalent. Using this equivalence and considering $d_{\min}(s,t)$, it is easily checked that $(\mathcal{N}_1, w_1)$ and $(\mathcal{N}, w)$ are equivalent.

Now assume that $\{s,t\}$ is a reticulated cherry in $(\mathcal{N}, w)$. Then $\{s,t\}$ is a reticulated cherry in $(\mathcal{N}_1, w_1)$ where, by Lemma 5.1, t is the reticulation leaf. Let $(\mathcal{N}_1', w_1')$ be the reticulation-pair weighted normal network with outgroup r obtained from $(\mathcal{N}_1, w_1)$ by isolating $\{s,t\}$. Since $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ are the minimum distance matrix and maximum distance outgroup vector of $(\mathcal{N}_1, w_1)$, it follows by Lemma 5.2 that $\mathcal{D}'$ and $\mathbf{d}'$ are the minimum distance matrix and maximum distance outgroup vector of $(\mathcal{N}_1', w_1')$. So, by the induction assumption, $(\mathcal{N}_1', w_1')$ and $(\mathcal{N}', w')$ are equivalent.

In $(\mathcal{N}, w)$, let p_s and p_t denote the parents of s and t, respectively, and let g_t denote the parent of p_t that is not p_s. Since $\mathcal{N}$ is normal, g_t is a tree vertex and not an ancestor of p_s. We next show that there is precisely one choice for the attachment of the edge in $(\mathcal{N}', w')$, and thus also in $(\mathcal{N}_1', w_1')$, corresponding to (g_t, p_t) in $(\mathcal{N}, w)$.

Since $\mathcal{N}$ is normal, there is a (directed) path P from g_t to a leaf, say ℓ, containing no reticulations. Since w is reticulation-pair, $d_{\min}(t,\ell)$ is the length of the up-down path whose union of edges consists of the edges in $\{(g_t, p_t), (p_t, t)\}$ and P. Thus, if we knew ℓ, then, to locate the place in $(\mathcal{N}', w')$ at which to insert g_t, we simply start at ℓ and follow the unique path against the direction of the edges towards the root until we reach a distance

$$d_{\min}(t,\ell) - (d_{\max}(r,t) - \mathcal{Q}_r(s,t))$$

from ℓ, since the bracketed term gives the combined length of (g_t, p_t) and (p_t, t). However, *a priori*, we do not know ℓ. So there are potentially $O(n)$ places in $(\mathcal{N}', w')$ at which we could insert g_t. We claim there is exactly one such place to insert g_t so that the resulting weighted network (after subdividing the edge incident with t, inserting a new vertex p_t and adding

the new edge (g_t, p_t)) has minimum distance matrix $\mathcal{D}_{\min}$ and maximum distance outgroup vector $\mathbf{d}_{\max}$ and no zero-length tree-edges.

We call a leaf ℓ' a *candidate leaf* if

- the path starting at ℓ' and going against the direction of the edges (and, thus, towards the root) a distance $d_{\min}(t, \ell') - (d_{\max}(r, t) - \mathcal{Q}_r(s, t))$ does not traverse a reticulation;
- the unique position along this path at a distance $d_{\min}(t, \ell') - (d_{\max}(r, t) - \mathcal{Q}_r(s, t))$ from ℓ', denoted $g_{\ell'}$, is not a vertex, that is, $g_{\ell'}$ is partway along an edge of $(\mathcal{N}', w')$; and
- $g_{\ell'}$ is not an ancestor of p_s.

Note that the unknown leaf ℓ is a candidate leaf. Moreover, if the second or third conditions were not satisfied and we tried to reconstruct a network by inserting g_t at position $g_{\ell'}$ we would either need to introduce zero-weight tree edges or we would introduce a shortcut, contradicting the assumptions about $(\mathcal{N}, w)$.

We now show that if $g_{\ell'}$ is not at the same position as g_t, then $g_{\ell'}$ is an ancestor of g_t. Suppose not, then a minimum length up-down path in $(\mathcal{N}, w)$ from t to ℓ' via g_t must traverse the edge containing position $g_{\ell'}$. But then the length of this up-down path is not $d_{\min}(t, \ell')$, by definition of $g_{\ell'}$. Likewise, a minimum length up-down path in $(\mathcal{N}, w)$ from t to ℓ' via p_s must traverse the edge containing position $g_{\ell'}$ (since p_s is not a descendant of $g_{\ell'}$). Again the length of this up-down path is not $d_{\min}(t, \ell')$. This contradicts the fact that $(\mathcal{N}, w)$ has minimum distance matrix $\mathcal{D}_{\min}$.

Finally, observe that if $g_{\ell'}$ is an ancestor of g_t, then the minimum path length between t and ℓ in the network obtained from $(\mathcal{N}', w')$ by adding an edge $(g_{\ell'}, p_t)$ will be strictly larger than $d_{\min}(t, \ell)$. This establishes the claim. Moreover, the correct position g_t can be found as the unique common descendant of all candidate positions $g_{\ell'}$. It now follows that, as $(\mathcal{N}_1', w_1')$ and $(\mathcal{N}', w')$ are equivalent, $(\mathcal{N}_1, w_1)$ and $(\mathcal{N}, w)$ are equivalent. This completes the proof of the uniqueness part of Theorem 2.4.

□

5.1. **The Algorithm.** Let $(\mathcal{N}, w)$ be a reticulation-pair weighted normal network on $X \cup \{r\}$, where r is an outgroup, and let $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ denote the minimum distance matrix and maximum distance outgroup vector of $(\mathcal{N}, w)$. The following algorithm, called Reticulation-Pair Normal, takes as input X, $\mathcal{D}_{\min}$, and $\mathbf{d}_{\max}$ and returns a weighted network $(\mathcal{N}_0, w_0)$ equivalent to $(\mathcal{N}, w)$ in which all reticulation edges have weight zero. As before, the proof of its correctness is essentially established in proving the

uniqueness part of the theorem and so is omitted. But its running time is given at the end.

1. If $|X| = 1$, then return the weighted phylogenetic network consisting of the single vertex in X.
2. If $|X| = 2$, say $X = \{r, s\}$, then return the phylogenetic network $(\mathcal{N}_0, w_0)$ consisting of leaves r and s adjoined to the root ρ with (ρ, r) and (ρ, s) positively weighted so that $d_{\max}(r, s) = w(\rho, r) + w(\rho, s)$.
3. Else, find a 2-element subset $\{s, t\}$ of X such that

$$\mathcal{Q}_r(s, t) = \max\{\mathcal{Q}_r(x, y) : x, y \in X\}.$$

   (a) If $d_{\min}(s, x) + d_{\max}(r, t) - d_{\max}(r, s) = d_{\min}(t, x)$ for all $x \in X$ (so $\{s, t\}$ is a cherry), then
      (i) Reduce t in $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ to give the distance matrix $\mathcal{D}'$ and distance vector $\mathbf{d}'$ on $X' = X - \{t\}$.
      (ii) Apply Reticulation-Pair Normal to input $X' \cup \{r\}$, $\mathcal{D}'$, and $\mathbf{d}'$. Construct $(\mathcal{N}_0, w_0)$ from the returned weighted phylogenetic network $(\mathcal{N}_0', w_0')$ on X' as follows. If p_s' is the parent of s in $(\mathcal{N}_0', w_0')$, then subdivide (p_s', s) with a new vertex p_s, adjoin a new leaf t to p_s via a new edge (p_s, t), and set

$$w_0(p_s, s) = d_{\max}(r, s) - \mathcal{Q}_r(s, t),$$

$$w_0(p_s', p_s) = w_0'(p_s', s) - w_0(p_s, s),$$

      and

$$w_0(p_s, t) = d_{\min}(s, t) - w_0(p_s, s).$$

      Keeping all other edges weight the same as their counterparts in $(\mathcal{N}_0', w_0')$, return $(\mathcal{N}_0, w_0)$.

   (b) Else ($\{s, t\}$ is a reticulated cherry, in which t is the reticulation leaf if there exists an $x \in X - \{s, t\}$ such that $d_{\min}(t, x) < d_{\min}(s, x) + d_{\max}(r, t) - d_{\max}(r, s)$),
      (i) Isolate $\{s, t\}$ in $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ to give the distance matrix $\mathcal{D}'$ and distance vector $\mathbf{d}'$ on X.
      (ii) Apply Reticulation-Pair Normal to input $X \cup \{r\}$, $\mathcal{D}'$, and $\mathbf{d}'$. Construct $(\mathcal{N}_0, w_0)$ from the returned weighted phylogenetic network $(\mathcal{N}_0', w_0')$ on $X \cup \{r\}$ as follows. For each leaf ℓ in $X - \{s, t\}$, follow the unique path starting at ℓ and going against the direction of the edges towards the root until either a reticulation or a distance

$$d_{\min}(t, \ell) - (d_{\max}(r, t) - \mathcal{Q}_r(s, t))$$

      from ℓ is reached. Amongst the points reached (which are not reticulations), insert a new vertex g_t in the unique point that is a descendant of all the other points reached and weight the edges incident with g_t appropriately. Now, subdivide the edge

incident with t with a new vertex p_t, add the new edge (g_t, p_t), and set $w_0(p_s, p_t) = 0$, $w_0(g_t, p_t) = 0$, and

$$w_0(p_t, t) = d_{\max}(r, t) - \mathcal{Q}_r(s, t),$$

where p_s is the parent of s. Keeping all other edges the same weight as their counterparts in $(\mathcal{N}_0', w_0')$, return $(\mathcal{N}_0, w_0)$.

For the running time, Reticulation-Pair Normal takes as input a set $X \cup \{r\}$, a $|X| \times |X|$ distance matrix $\mathcal{D}_{\min}$ whose entries are the minimum-length of an up-down path joining elements in X, and a distance vector $\mathbf{d}_{\max}$ of length $|X|$ whose entries are the maximum length of an up-down path from r to each element in X of a reticulation-pair normal network $(\mathcal{N}, w)$ on $X \cup \{r\}$, where r is an outgroup. If $|X| \in \{1, 2\}$, then the algorithm runs in constant time. If $|X| \geq 3$, each iteration begins by finding a 2-element subset $\{s, t\}$ of X such that

$$\mathcal{Q}_r(s, t) = \max\{\mathcal{Q}_r(x, y) : x, y \in X\}.$$

This takes $O(|X|^2)$ time, and once such a 2-element subset is found, we construct $\mathcal{D}'$ and $\mathbf{d}'$. This construction is done in one of two ways depending on whether or not

$$d_{\min}(s, x) + d_{\max}(r, t) - d_{\max}(r, s) = d_{\min}(t, x)$$

for all $x \in X - \{s, t\}$. If, for some x,

$$d_{\min}(s, x) + d_{\max}(r, t) - d_{\max}(r, s) \neq d_{\min}(t, x),$$

we need to additionally check which of

$$d_{\min}(s, x) + d_{\max}(r, t) - d_{\max}(r, s) < d_{\min}(t, x)$$

and

$$d_{\min}(s, x) + d_{\max}(r, t) - d_{\max}(r, s) > d_{\min}(t, x)$$

holds. Thus the determination of which way to construct $\mathcal{D}'$ and $\mathbf{d}'$ can be done in $O(|X|)$ time. Whether $\mathcal{D}'$ and $\mathbf{d}'$ is constructed by reducing an element of X or isolating a 2-element subset of X, the construction can be done in $O(|X|)$ time. Once $(\mathcal{N}_0', w_0')$ is returned, it takes constant time to augment to $(\mathcal{N}_0, w_0)$ if $\mathcal{D}'$ and $\mathbf{d}'$ have been obtained from $\mathcal{D}_{\min}$ and $\mathbf{d}_{\max}$ by reducing. Otherwise, we need to find the unique place in $(\mathcal{N}_0', w_0')$ to insert g_t. Since $(\mathcal{N}, w)$ has $O(|X|)$ vertices in total [2], it takes at most $O(|X|^2)$ time to find the possible locations in which to insert g_t. Finding the correct location, the one that is a descendant of all the others, can be done in $O(|X|)$ time by repeatedly deleting vertices of out-degree zero until the first possible location appears as a vertex of out-degree zero. Thus $(\mathcal{N}_0, w_0)$ can be returned in $O(|X|^2)$ time, and so the total time of each iteration is $O(|X|^2)$.

When we recurse, the distance matrix $\mathcal{D}'$ and distance vector $\mathbf{d}'$ inputted to the recursive call is the minimum distance matrix and maximum distance

outgroup vector of a normal network with either one less leaf or one less reticulation than $(\mathcal{N}, w)$. As normal networks have at most $|X|-2$ reticulations, and therefore $O(|X|)$ vertices in total [2], the total number of iterations is at most $O(|X|)$. Hence Reticulation-Pair Normal completes in $O(|X|^3)$ time, thereby completing the proof of Theorem 2.4.

6. Conclusion

In this paper, we established two analogues of the Tree-Metric Theorem for normal networks. The first analogue, Theorem 2.3, shows that a normal network with an equidistant weighting is determined, up to a certain equivalence, by the minimum-lengths of the up-down paths joining pairs of taxa. The second analogue, Theorem 2.4, shows that a normal network with an outgroup r and reticulation-pair weighting is determined, up to a certain equivalence, by the minimum-lengths of the up-down paths joining pairs of taxa as well as the maximum-lengths of the up-down paths joining r and a single taxa. Previously, these results were established by using the lengths of all up-down paths joining pairs of taxa of which there could be exponentially-many such lengths.

Normal networks are a rich class of phylogenetic networks. Knowing that a single measure of distance between each pair of taxa is sufficient to determine such networks with certain weightings, we are now interested in developing a practical algorithm for constructing normal phylogenetic networks from distances. On the biological side, this will require the development of appropriate models to take into account that we aim to reconstruct networks from their minimal distances. On the algorithmic side, the main issue will be to develop approaches to deal with noisy data. In this regards, the distance-based, tree-building algorithm Neighbor Joining [11] may provide some clues on how to proceed. More specifically, the Neighbor Joining algorithm is based on recursively constructing trees by picking pairs of taxa which correspond to cherries in case the input distance is a tree-distance. Thus, to build normal networks from biological distances, one approach could be to develop an algorithm based on recursively picking pairs of taxa corresponding to cherries and reticulated cherries.

Acknowledgements

We thank the anonymous referees for their helpful comments. Katharina Huber and Vincent Moulton also thank the Biomathematics Research Centre at the University of Canterbury for its hospitality.

References

[1] Baroni M, Semple C, Steel, M (2006) Hybrids in real time. Syst Biol 55: 46–56

[2] Bickner DR (2012) On normal networks. PhD thesis, Iowa State University, Ames, Iowa

[3] Bordewich M, Semple, C (2016) Determining phylogenetic networks from inter-taxa distances. J Math Biol 73: 283–303

[4] Bordewich M, Semple C, Tokac N Constructing tree-child networks from distance matrices. Algorithmica, in press

[5] Bordewich M, Tokac N (2016) An algorithm for reconstructing ultrametric tree-child networks from inter-taxa distances. Discrete Appl Math 213: 47–59

[6] Buneman, P (1974) A note on the metric property. J Combin Theory Ser B 17: 48–50

[7] Cardona G, Rossello F, Valiente, G (2009) Comparison of tree-child networks. IEEE/ACM Trans Comput Biol Bioinformatics 6: 552–569

[8] Chan H-W, Jansson J, Lam T-W, Yiu S-M (2006) Reconstructing an ultrametic galled phylogenetic network from distance matrix. J Bioinform Comput Biol 4: 807–832

[9] Huber KT, Scholz GE Beyond representing orthology relationships by trees. Algorithmica, in press

[10] McDiarmid C, Semple C, Welsh D (2015) Counting phylogenetic networks. Ann Comb 19: 205–224

[11] Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425

[12] Semple C, Steel M (2003) Phylogenetics. Oxford University Press

[13] Simões-Pereira JMS (1969) A note on the tree realization of a distance matrix. J Combin Theory 6: 303–310

[14] Sokal RR (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38: 1409–1438

[15] Willson SJ (2007) Unique determination of some homoplasies at hybridization events. Bull Math Biol 69: 1709–1725

[16] Willson SJ (2007) Reconstruction of some hybrid phylogenetic networks with homoplasies from distances. Bull Math Biol 69: 2561–2590

[17] Willson SJ (2010) Properties of normal phylogenetic networks. Bull Math Biol 72: 340–358

[18] Willson SJ (2012) Tree-average distances on certain phylogenetic networks have their weights uniquely determined. Algorithm Mol Biol 7: 13

[19] Yu Y, Nakleh L (2015) A distance-based method for inferring phylogenetic networks in the presence of incomplete lineage sorting. In: Harrison R et al (eds) International Symposium on Bioinformatics Research and Applications (ISBRA), LNBI 9096, pp 378–389

[20] Zaretskii KA (1965) Constructing trees from the set of distances between vertices. Uspehi Matematiceskih Nauk 20: 90–92

Department of Computer Science, Durham University, Durham, DH1 3LE, United Kingdom

E-mail address: m.j.r.bordewich@durham.ac.uk

School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom

E-mail address: k.huber@uea.ac.uk

School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom

E-mail address: v.moulton@uea.ac.uk

School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: charles.semple@canterbury.ac.nz